Review and analysis of changes in prolonged exposure of various rat tissues to space conditions

Nikoloska Nora Kotevski Stefan Tasevski Mladen

{nora.nikoloska,stefan.kotevski,mladen.tasevski}@students.finki.ukim.mk

June 2020

Abstract

In this paper we look at different rat tissues affected by prolonged stay on the International Space Stations and space flight. Some expected effect include muscle atrophy, sight impairment, bone decay and decreased immunity. Some of these effects have been show to appear in astronauts as well. The NASA GeneLab missions are designed in order to pinpoint the changes in the gene expression in different species aboard the International Space Station. In this project we compare the extracted RNA from the RR-6 mission in order to calculate differential genetic expression among the collected samples. We use this data to define possible fluctuations in the gene expression of species subject to space, space flight and extraterrestrial environment and the changes experienced as a result . . .

1 Prepossessing

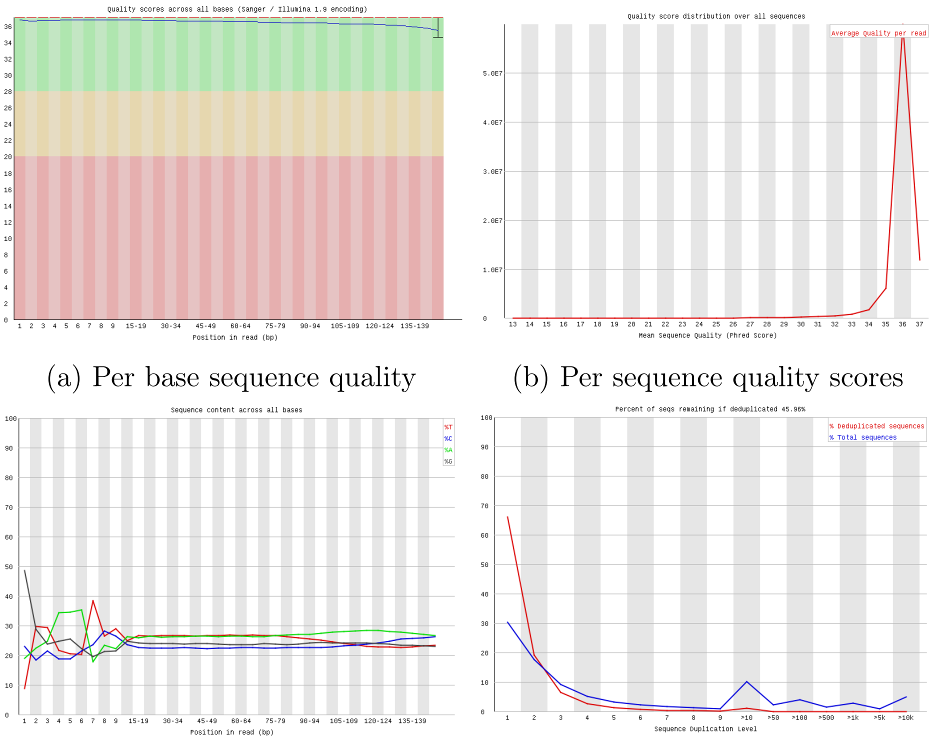
The data used in this project is publicly available on the GeneLab website. Here we give a broad overview of the prepossessing stage. First the data quality is checked as illustrated in the following sections. After a quality check is preformed the data can be processed in order to ensure that all reads meet the standards. This allows us to insure that only important parts are kept and analysed and the noise is filtered out. Next Kallisto bustools are used for the alignment because of the highly optimized pseudo-alignment method. This method is proven to give a substantial processing time decrease compared to modern aware alignment methods. From this a count matrix can be generated and the result is used to visualize the gene expression of the samples and find the over-expressed genes.

1.1 FastQ format

In the first step we choose one of the recent rodent missions (RR-6) from which we download the data.[4]. The data is formated in the FastQ format which includes a quality rating for each base in the sequence. This is obtained by using ASCII encoding for the quality values starting from position 33.

1.2 FastQC Report and analysis

The pipeline fist inspects the downloaded raw FastQ file in order to generate a FastQC report. From the generated report it can be concluded that there are no reads with poor quality and the “Per sequence GC content” and sequence length distribution tests are successful. However, the “Per base sequence content” test failed due to fluctuations in the base count in the initial positions. This effect is considered to be expected and should not trimmed. In addition, there is a high percentage of duplicated and over-represented sequences detected. However, there are no matches against the Illumina adapters which is expected due to the nature of the experiments. The counts of the duplicates are needed to conclude the gene expression. We are therefore proceeding without duplicates removal. For the trimming of adapters we experimented with two tools. Trimomatic a sequence trimming tool was used but since the FastQC report had not detect any adapter matches nothing could be referenced. A more sucessfull aproach was achieved using TrimGalore. This tool was able to trim the adapters and improve the dataset quality substantially since it is a tool officially supported by NASA-Genelab. The results can be seen in **Figure 2**.



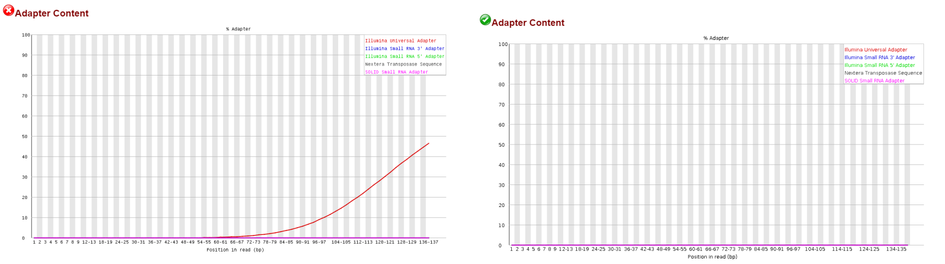
(a) Per base sequence quality

(b) Per sequence quality scores

(c) Per base sequence content

(d) Sequence duplication levels

Figure 1: Graphs from FastQC Report and analysis



(a) Before trimming

(b) After trimming with TrimGalore

Figure 2: Adapter Content graph from FastQC Report and analysis

1.3 Kallisto pseudo-alignment

The tools we use in the pipeline for alignment and analysis of the reads are Kallisto bustools programs[2]. Kallisto offers a significant speed-up in the alignment process, by implementing the concept of pseudo-alignment[1].

The figure represents a de-Bruijn graph (T-DBG) of the transciptome. The sequence is transformed in multiple k-mers represented by the circles. In the graph, a circle is colored in black if the k-mer is aligned with the corresponding transcript. If several k-mers in a row are compatible with the same set of transcripts, the graph is simplified by adding the dots.

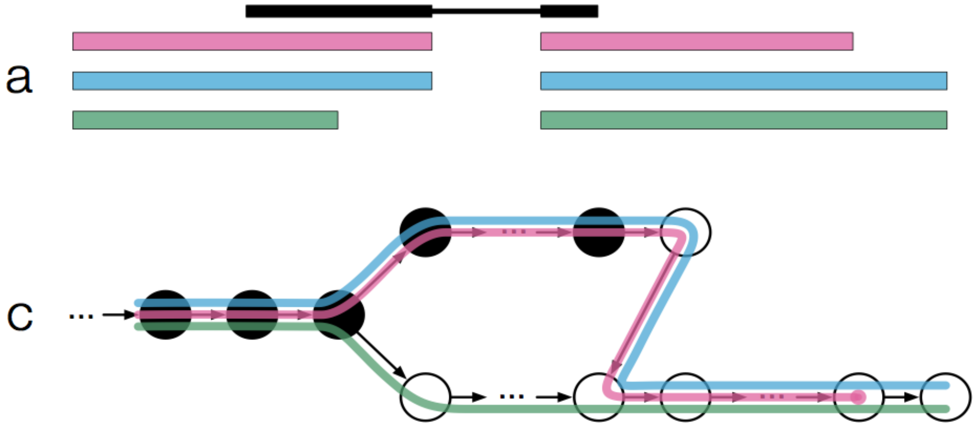


Figure 3: The Kallisto pipeline

The first step in the process is downloading the publicly available mouse genome against which the reads are to be aligned. Next, we are creating indexes on the genome and specifying two streams of data from tbe GLDS-248 study. The streams are the corresponding paired reads of one sample. The data is processed by using the `bus` command.

1.4 Generating count matrix with bustools

Upon completion of the process of alignment, we use bustools as the next tool in the pipeline in order to generate a count matrix. The inputs in the *correct* command are the *.bus* file generated with kallisto, a mapper file of genes to transcripts for the given genome and the white-list file. Upon completion of the process three files are created. Two files for the genes and transcripts respectfully and the count matrix in a *.mtx* format.

1.5 Identification of highly variable genes

In the next step of the process the data is converted into an AnnData object with the transcript as observations (obs) and the genes as variables (vars). Fist, we filter the genes to only consider ones with more than one count and additional basic filtering is performed. Then, we normalize the data by using the normalize_per_cell function and we pass an optional argument to only consider cells with more than one count. By using the command `sc.pp.highly_variable_genes` we are visualise the highly expressed genes. Afterwards, we log transform the data and scale according to the data mean and variance. We also check the count number for some genes after the performed normalisation.

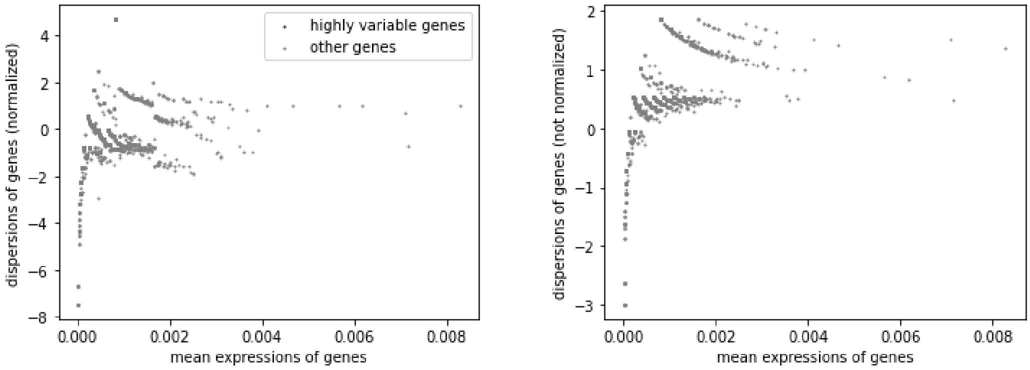


Figure 4: Highly variable genes

2 Differential gene expression analysis

The next challenge is to compute the genes with the highest differential expression. We compare these genes across samples in order to detect the genes whose expression is altered the most in space environment. By using the function highest_exprchar_genes we plot the genes with the highest expression.

2.1 Computing neighborhood graph and Louvain clustering

The next step of the pipeline is to compute the neighbourhood graph by using the neighbours function. In order to cluster the genes we use the Louvain[3] algorithm on the neighbourhood graph. This method is a greedy optimization method that aims to detect communities in a network and form

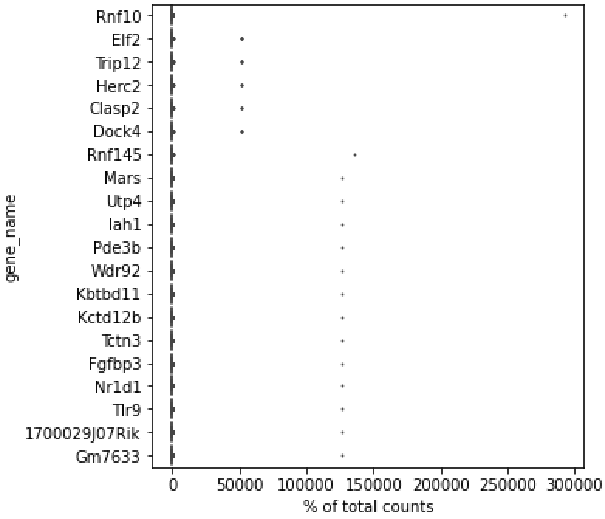


Figure 5: Genes with highest expression

clusters. We experimented with different values for the resolution parameter which result in a different number of clusters. Using a resolution value of 0.08 the algorithm yielded 9 clusters from which we proceed with the marker genes identification.

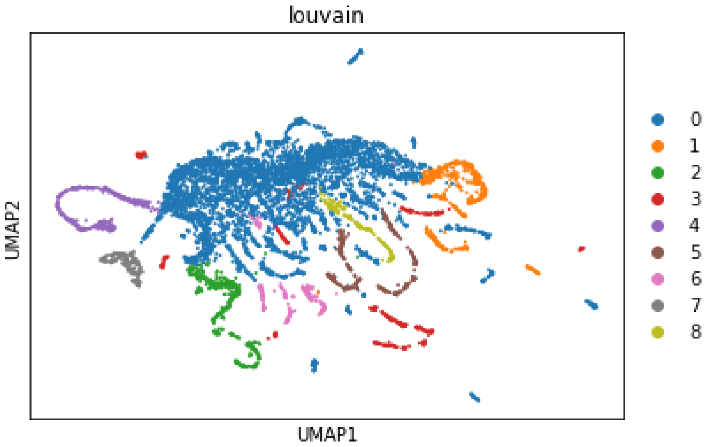
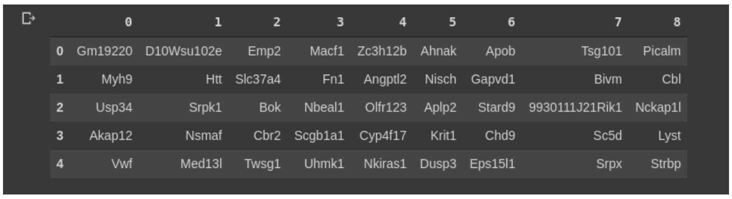


Figure 6: Visualisation of the Louvain clusters

2.2 Identification of marker genes

We use the top genes of each cluster for the upcoming comparison across different samples. For each experiment we will use the samples from mice aboard the ISS. The function of the genes with the highest frequency across all analyzed samples will be determined using gene ontology.



	0	1	2	3	4	5	6	7	8
0	Gm19220	D10Wsu102e	Emp2	Macf1	Zc3h12b	Ahnak	Apob	Tsg101	Picalm
1	Myh9	Htt	Slc37a4	Fn1	Angptl2	Nisch	Gapvd1	Bivm	Cbl
2	Usp34	Srpk1	Bok	Nbeal1	Olfr123	Aplp2	Stard9	9930111J21Rik1	Nckap1l
3	Akap12	Nsmaf	Cbr2	Scgb1a1	Cyp4f17	Krit1	Chd9	Sc5d	Lyst
4	Vwf	Med13l	Twsg1	Uhmk1	Nkiras1	Dusp3	Eps15l1	Srpx	Strbp

Figure 7: Extraction of the 5 top genes by cluster

3 Comparison of marker genes across samples

In order to understand the impact of spaceflight on the overall mouse phenotype, we tracked the marker genes across samples of four different tissues. As part of the Rodent Research-6 (RR6) Genelab Mission, samples from lung[4], colon[5], thymus[6] and dorsal skin[7] tissue were collected. For each tissue type, we recorded the marker genes for 3 different samples, thus yielding 12 experiments with different datasets. The goal is to depict the most common genes that appear as marker genes across the samples and analyze their function. In the results, we detected 2 genes that occur 8 out of 12 times as a marker genes, 2 genes whose occurrence is 7 times and 3 genes that occur 6 times and 5 genes that occur 5 times. We use gene ontology to continue with the analysis of the function of the top four genes: Macf1[8], Neb[9], Ahnak[10], D10Wsu102e[11]. We use the Ensembl[12] gene ontology database to search the genes by their ENSMUS name, and the Mouse Genome Informatics databse for gene function.

3.1 Macf1

This gene's functions in the phenotype are detected in embryo development, cellular level, metabolism, integument, mortality and or aging, nervous system, respiratory system and vision.

3.2 Neb

The systems this gene is associated with are adipose tissue, neurological, growth, metabolism, mortality and or aging, muscle, skeleton and vision. As this gene is essential for growth, the inhibition of the gene in the process of embryo development can be lethal.

3.3 Ahnak

This gene is expressed in cells in the adipose tissue, has a cellular function and influences the growth, metabolism and mortality/aging as well as the hematopoietic and immune system.

3.4 D10Wsu102e

Due to the expression of the gene in almost every aspect of the phenotype and it's broad function on the organism, gene function on concrete systems cannot be defined.

4 Review and conclusion

Space medicine and space microbiology are vast research fields that aim to understand the effects and dangers that humans face in space environment and provide suitable solutions. However, by using rodents, researchers are able to track the changes in the organisms more quickly and accurately, without causing safety issues on human health. Many studies agree that space environment affects the cardiovascular, musculoskeletal, respiratory system and vision. As discussed in the last section the genes that occur the most as deferentially expressed genes in the rodent studies are functioning as part of these systems. Some of the human pathophysiological adaptations include[13]: reduced sight, cases of muscle atrophy and bone damage and significant adaptations in the respiratory system and ventilation. Many of the mentioned side-effects of space flight and space environment exposure in turn cause other indirect effects such as: effecting the reflex mechanisms and the endocrine system which can cause body weight fluctuations, excessive urination and sodium excretion. The increased number of studies and data repositories are beneficial in the process of further understanding of the impact of prolonged exposure to a space environment on live organisms. This can eventually lead to the development of new technologies and techniques for conquering space and starting the colonization of the Solar System.

References

[1] Nicolas L. Bray, Harold Pimentel, Páll Melsted and Lior Pachter. *Near-optimal RNA-Seq quantification*

[2] ProfilePáll Melsted, Sina Booeshaghi, Fan Gao, Eduardo Beltrame, Lambda Lu, ProfileKristján, Eldjárn Hjorleifsson, ProfileJase Gehring, ProfileLior Pachter. *Modular and efficient pre-processing of single-cell RNA-seq*

[3] Itamar Kanter, Gur Yaari, Tomer Kalisky *Applications of community detection algorithms to large biological datasets*

[4] NASA GeneLab. *GLDS-248: Transcriptional analysis of lung from mice flown on the RR-6 mission*

[5] NASA GeneLab. *GLDS-247: Transcriptional analysis of colon from mice flown on the RR-6 mission*

[6] NASA GeneLab. *GLDS-244: Transcriptional analysis of thymus from mice flown on the RR-6 mission*

[7] NASA GeneLab. *GLDS-243: Transcriptional analysis of dorsal skin from mice flown on the RR-6 mission*

[8] Mouse Genome Informatics *Macf1 - Gene Detail*

[9] Mouse Genome Informatics *Neb - Gene Detail*

[10] Mouse Genome Informatics *Ahnak - Gene Detail*

[11] Mouse Genome Informatics *D10Wsu102e - Gene Detail*

[12] Ensembl genome browser *Ensembl*

[13] Frontiers in Psychology *Human Pathophysiological Adaptations to the Space Environment*